DOCUMENT RESUME

ED 249 256 TM 840 599

AUTHOR McGuire, Dennis P.
TITLE Methodological Considerations When Using the 1979-80 Reading/Literature Public Use Tapes: Data Analysis, Identification of Traits, and Construction of Scales from NAEP Items.
SPONS AGENCY Minnesota Univ., Minneapolis.
PUB DATE Apr 84
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984). Portions may be marginally legible.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; *Computer Software; *Educational Assessment; Elementary Secondary Education; Federal Programs; Literature; Reading Achievement; *Research Methodology; Research Problems; *Statistical Analysis; Test Construction; Testing Programs; Test Reliability; Test Results
IDENTIFIERS *National Assessment of Educational Progress; Second Literature Third Reading Assessment (1980); *Statistical Package for the Social Sciences

ABSTRACT

Efficient methods of using the Statistical Package for the Social Sciences (SPSS) to analyze National Assessment of Educational Progress (NAEP) data files are discussed. One error in the NAEP SPSS file is discussed, and another error (which may be system-dependent) is mentioned. In addition, purely mathematical methods are used to address the questions of whether there is evidence for the existence of four hypothesized traits (Words and Phrases, Lexical Relationships, Propositional Relationships, and Textual Relationships). Further, the question of estimating reliability of a new set of items gathered from different NAEP booklets is addressed. It is concluded that the current data do not support the thesis that these traits are distinct, although there is some evidence that Words & Phrases and Lexical Relationships are distinct. It is further concluded that the reliability of new sets of items cannot be safely estimated at this time. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED249256

Methodological Considerations When Using the 1979-80 Reading/Literature Public Use Tapes: Data Analysis, Identification of Traits, and Construction of Scales from NAEP Items

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Dennis P. McGuire

Department of Educational Psychology

University of Minnesota

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D.P. McGuire

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented to the 1984 annual meeting of the American Educational Research Association, New Orleans, April 25, 1984.

TM 840 595

Note:

This research was funded, in part, by a Dissertation Fellowship from the Graduate School at the University of Minnesota and by a grant from the University of Minnesota Computer Center.

Copyright (c) 1984 by Dennis P. McGuire

Abstract

Efficient methods of using SPSS to analyze NAEP data files are discussed. One error in the NAEP SPSS file is discussed, and another error (which may be system-dependent) is mentioned. In addition, purely mathematical methods are used to address the questions of whether there is evidence for the existence of four hypothesized traits (Words and Phrases, Lexical Relationships, Propositional Relationships, and Textual Relationships). Further, the question of estimating reliability of a new set of items gathered from different NAEP booklets is addressed. It is concluded that the current data do not support the thesis that these traits are distinct, although there is some evidence that Words & Phrases and Lexical Relationships are distinct. It is further concluded that the reliability of new sets of items cannot be safely estimated at this time.

Table of Contents

Abstract 2

I. Computer Analyses: Saving Time and Money . . . 4

Discussion 5
Figure 1: Error in SPSS "COMMENT" cards . . . 8
Figure 2: SPSS Error: Constructed Format Too Long 10

II. Identification of Traits 12

Rationale 12
Table 1: Scale Reliability Estimated From Mean Interitem Correlation 14
Methodology 16
Results 18
Table 2: Mean Item Intercorrelations For All Items 19
Table 3: Summary Statistics of "Inappropriate" Items 23
Table 4: Mean Item Intercorrelations After Deleting "Inappropriate" Items . . . 25
Conclusions 26

III. Combining Items from Different Booklets 27

Rationale 28
Methodology 31
Results 33
Conclusions 35

Bibliography 38

Appendix 39

I. Computer analyses: Saving time and money

The National Assessment of Educational Progress provides lovely data tapes: not only are the data provided but SAS, SPSS and Osiris files are included to make it easier to analyze the files. All the analyst need do is read the appropriate files from the NAEP tapes, run the analyses, and save the output. Analyses performed for this study were done using SPSS, the Statistical Package for the Social Sciences (Nie, Hull, Jenkins, Steinbrenner and Bent, 1975). Unfortunately, there are errors and inconveniences in peforming analyses. Some of them result from the design of the NAEP study, while the others can be corrected or worked around as necessary. First, it is necessary to compute the proper weighting factor (which changes depending on the analysis needed). For the analyses discussed here, it was then necessary to recode the relevant cognitive variables to right-wrong scoring, using the COMPUTE and IF cards included as COMMENT cards within the SPSS file. Those cards contain an error. Finally, as the analyses proceeded, it became clear that the most direct ("easiest") way to do an analysis was often the costliest. For instance, as NAEP points out (NAEP, 1981, p. 45) if the examples presented are followed, minimal programmer time is needed but computer costs are greater. As experience was gained in performing analyses,

easy ways to save both programmer time and computer costs were discovered.

McConeghy and Miller (1983) have discussed ways to plan a study to reduce both programmer time and computer costs. Both NAEP (1981) and McConeghy and Miller (1983) discuss methods to create SPSS system files utilizing items from different booklets. The following discussion adds a few ideas about saving time and money on analyses. A different (but so far unsuccessful) approach to combining items from different booklets is discussed in Part III.

Discussion

The analyses discussed here used only the scores from certain NAEP cognitive items and the WEIGHTS variable. The first step was to compute the mean of WEIGHTS so that further analyses could be done with proper subject weighting. It cost roughly $12.50 in computer time to compute the mean for WEIGHTS for each NAEP booklet using the full SPSS file provided by NAEP. However, only a few SPSS cards are really needed, as shown by this example for Age 9 Booklet 5 analyses:

```
RUN NAME        WEIGHTS RUN: AGE 09 BOOKLET 05
DATA LIST       FIXED/1
                WEIGHTS 65-73(2)
N OF CASES      2627
CONDESCRIPTIVE WEIGHTS
STATISTICS      1
FINISH
```

These SPSS cards can be run on the Control Data Cyber 825 computer at the University of Minnesota by SUBMITting the following job file. The output is saved on disk and can be printed later, if desired. Since printing costs can be a major expense ($.033 per page at Minnesota), it is worthwhile minimizing unneeded output.

```
/JOB
NAEP,T100.
/USER,account #.
UNPAGE.
A,DATAN05=data file,SPSSN05=SPSS program.
SPSS,D=DATAN05,I=SPSSN05,RL=400.
UNPAGE.
RETAIN,OUTPUT=OUTN05.
```

Using this abbreviated set of SPSS cards saved 80% to 90% of the cost of a WEIGHTS run. Within each age group, the WEIGHTS variable consistently was in the same position (columns 65-73 here), so the analyses are particularly quick when several booklets are analyzed: few changes need to be made in the SPSS deck to perform an analysis on a new booklet. NAEP could have printed out the mean of WEIGHTS for each booklet, but did not, perhaps because different types of analysis require different weighting variables (see NAEP, 1981, pp. 28ff). The values

obtained in this study are presented in Appendix I with the hope that they will save future researchers some time and money.

The lesson to be learned here is that eliminating unnecessary variables can save a lot of money. Any substantive analysis will require more than one variable, so cost savings will be smaller, but still could be sizeable.

NAEP provides a set of COMMENT cards in the SPSS program that can be used to recode the cognitive items to right-wrong scoring. An error in the SPSS deck was discovered that had to be corrected in each booklet (See Figure 1). The problem is that in the MISSING VALUES and VALUE LABELS cards for the recoded variables, variable Y0101A is identified (for age 17; W0101A and X0101A are identified for ages 9 and 13, respectively) and yet that variable does not exist. To correct the problem, it is necessary to check the first COMPUTE card to find the name of the first recoded variable, and then change the erroneous "0101A" variable name to the correct name. That variable name changes from booklet to booklet, although most are "nn03A", where "nn" is the booklet number. After this change is made, the Age 9 and Age 13 analyses can be done.

There was an additional problem with the Age 17 SPSS programs that was almost impossible to interpret. I finally telephoned SPSS in Chicago, who referred me to Philip Galanter

Figure 1

Error in NAEP set of SPSS "COMMENT" cards

```
          (4) NO OPINION (8) NO RESPONSE (9) MISSING VALUE/
S0818G    (1) VERY HELPFUL (2) HELPFUL (3) NOT HELPFUL
          (4) NO OPINION (8) NO RESPONSE (9) MISSING VALUE/
S0818H    (1) VERY HELPFUL (2) HELPFUL (3) NOT HELPFUL
          (4) NO OPINION (8) NO RESPONSE (9) MISSING VALUE/
COMMENT
COMMENT                SCORING THE EXERCISES
COMMENT
COMMENT        THE FOLLOWING COMMENT CARDS MAY BE EDITED TO
COMMENT        CREATE SCORED VARIABLES WITH VARIABLE
COMMENT        LABELS, VALUE LABELS AND MISSING VALUE
COMMENT        ASSIGNMENTS.  VARIABLES ARE CREATED FOR ALL
COMMENT        EXERCISES WITH CORRECT FOILS.  THE CORRECT
COMMENT        FOIL FOR AN EXERCISE BECOMES 1, A MISSING
COMMENT        VALUE BECOMES A MISSING VALUE OF 9 AND ALL
COMMENT        OTHER RESPONSES BECOME 0.  THUS THE I DON'T
COMMENT        KNOW OR NO RESPONSE FOILS WOULD BE INCORRECT
COMMENT        RESPONSES.  THE VARIABLE NAME REMAINS THE
COMMENT        SAME WITH THE FIRST CHARACTER N, T, S OR A
COMMENT        CORRESPONDING TO AGE 09, 13, 17 OR ADULT
COMMENT        MODIFIED TO W, X, Y OR Z RESPECTIVELY.
COMMENT
COMMENT        REMOVE THE FIRST FIFTEEN CHARACTERS TO MAKE
COMMENT        THE FOLLOWING COMMENT CARDS OPERATIONAL.
COMMENT
COMMENT        COMPUTE         Y0803A=0
COMMENT        IF              (S0803A=4) Y0803A=1
COMMENT        IF              (S0803A=9) Y0803A=9
COMMENT        COMPUTE         Y0803B=0
COMMENT        IF              (S0803B=3) Y0803B=1
COMMENT        IF              (S0803B=9) Y0803B=9
COMMENT        COMPUTE         Y0803C=0
COMMENT        IF              (S0803C=3) Y0803C=1
               ...             ...

COMMENT        IF              (S0817G=1) Y0817G=1
COMMENT        IF              (S0817G=9) Y0817G=9
COMMENT        COMPUTE         Y0817H=0
COMMENT        IF              (S0817H=1) Y0817H=1
COMMENT        IF              (S0817H=9) Y0817H=9
COMMENT        MISSING VALUES
COMMENT                        Y0101A TO Y0817H(9)/
COMMENT        VALUE LABELS
COMMENT                        Y0101A TO Y0817H (0) INCORRECT
COMMENT                          (1) CORRECT (9) MISSING VALUE/
COMMENT        VAR LABELS
COMMENT                        Y0803A   H-663000 PHONE DIR:STORES SELL MILK LIST/
COMMENT                        Y0803B   H-663000 PHONE DIR:HENDRICKS MINING ON 6/
COMMENT                        Y0803C   H-663000 PHONE DIR:STAR TRACKER OPEN TO /
COMMENT                        Y0805B   H-306000 DWELLERS:CITY CHILDREN BETTER-C/
COMMENT                        Y0805C   H-306000 DWELLERS:CITY BETTER MENTALLY-M/
COMMENT                        Y0805D   H-306000 DWELLERS:MAIN PURPOSE-CITY AS H/
COMMENT                        Y0806A   H-822000 WRITING TYPE2:SHORT STORY      /
COMMENT                        Y0806B   H-822000 WRITING TYPE2:PERSONAL LETTER  /
COMMENT                        Y0806C   H-822000 WRITING TYPE2:FAIRY TALE       /
COMMENT                        Y0807A   H-801000 EMOT RESP:EVER LAUGHED WHEN REA/
COMMENT                        Y0807B   H-801000 EMOT RESP:EVER FELT BADLY WITH/
```

at Northwestern University's Vogelback Computer Center. He was able to explain the problem. Figure 2 shows a part of a printout with "ERROR NUMBER 147: DATA LIST - CONSTRUCTED FORMAT TOO LONG - 4000 CHARACTERS". This may be a machine-dependent problem, occurring only on Control Data Cyber computers, and has a straightforward explanation: when a DATA LIST is used, SPSS allots a 4000-character matrix to hold format information on the variables. The matrix can hold many variables, but the DATA LIST does not necessarily fill it efficiently. In this case, the 4000 character limit was exceeded in the analysis and the SPSS program could not run. Nothing can be done to correct this error, although it is possible to work around the problem. The solution is to eliminate a large number of variables and then run a reduced set of variables in the analysis. This was fine for our work, but if a full system file is needed for a project, it must be created as two or more small files which can then be merged into a full system file. Alternately, the DATA LIST can be modified or changed to a VARIABLE LIST - INPUT FORMAT combination in an attempt to run a single analysis. This is probably more trouble than it is worth.

In general, the more you can trust yourself with computer analyses, the more money you can save with them. If a file can be edited quickly to remove unneeded variables, then a dramatic reduction in computing costs can result. This is strongly

Figure 2

SPSS Error: Constructed Format too Long

```
UNIVERSITY COMPUTER CENTER
UNIVERSITY OF MINNESOTA

S P S S - - STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES

VERSION 8.3 (NOS) -- JANUARY 21, 1982

377700 CM MAXIMUM FIELD LENGTH REQUEST


          RUN NAME      AGE 17 BOOKLET 08 WEIGHTS RUN
          TASK NAME     NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS
          DATA LIST     FIXED/1
                        ADMNDATE 6-11 (0)
                        ADMNTIME 12-15 (0)
                        ADVPLENG 334-334 (0)
                        AGE      1-1 (0)
                        AGE09US  138-138 (0)
                        AGE13US  137-137 (0)
                        AVAIDS   297-297 (0)

                        TEACHRCH 299-299 (0)
                        THEATER  322-322 (0)
                        TIMEREAD 390-391 (0)
                        TOC      67-67 (0)
                        TVWATCHD 392-393 (0)

THE FOLLOWING HAS CAUSED AN ERROR.  TVWATCHD  (OCTAL 2426270124031004555S)

ERROR NUMBER.  147. PROCESSING CEASES, ERROR SCAN CONTINUES.

                        TWOCARS  394-394 (0)
                        TYPEWRTR 395-395 (0)
                        VACCLNR  396-396 (0)
                        VOCDEVEL 326-326 (0)
                        WEIGHTS  68-76 (2)
                        S0801A   401-401 (0)
                        S0801B   402-402 (0)
                        S0802A   403-403 (0)

    - -   - -  ERROR SUMMARY  - - - - -


    ERROR NUMBER.  51
    A VARIABLE LIST ON THE VALUE LABEL OR VAR LABEL CARD IS INVALID

    ERROR NUMBER  147
    DATA LIST - CONSTRUCTED FORMAT TOO LONG - 4000 CHARACTERS

    ERROR NUMBER.  192
    A VARIABLE LIST ON THE MISSING VALUE CARD IS INVALID

TOTAL CPU TIME USED .       1.996 SECONDS

RUN COMPLETED

NUMBER OF CONTROL CARDS READ 582
NUMBER OF ERRORS DETECTED     41
```

encouraged whenever possible. If one of the NAEP SPSS files is to be edited, it will be easiest to print out a copy of the entire file and then use the hard copy as a guide for editing. If multiple SPSS files are to be edited, it may be adequate to only print out one file in each age group involved since the files for different booklets are so similar. Researchers planning to combine data from different booklets should examine both the NAEP (1981) and McConeghy and Miller (1983) recommendations.

II. Identification of Traits

This section describes an attempt to address the specific question, "Do the four scales of NAEP Classification 23 (Words and Phrases, Lexical Relationships, Propositional Relationships and Textual Relationships) constitute distinct traits?" This question was approached through mean interitem correlations rather than reliability coefficients since mean interitem correlations can be directly compared while reliability estimates change with the number of items in a scale.

Rationale

Observed scale reliabilities depend on the number of items in the scale: the more items, the higher the reliability. Gulliksen (1950, p. 85) recommends comparing scale reliabilities through the scale-length invariant quantity $n[(1/alpha)-1]$, where n is the number of items in the scale and alpha is the observed reliability. This quantity equals $[(1/r)-1]$, where r is the mean interitem correlation (Actually, Gulliksen posed the problem differently, but it is re-stated here to fit this context more closely). Cronbach (1951, p. 323) made a suggestion that provides more interpretable

results: estimate mean interitem correlation from the observed reliability, $\underline{r}$=alpha/[$\underline{n}$+(1-$\underline{n}$)alpha], and compare scales by comparing mean interitem correlations. This formula for $\underline{r}$ is the inversion of the generalized Spearman-Brown formula (Gulliksen, 1950, p. 78; Lord and Novick, 1968, p. 90), and is better than Gulliksen's suggestion because a larger mean correlation relates directly to a higher scale reliability. Table 1 is a table of estimated scale reliabilities resulting from various combinations of mean interitem correlation and number of items in the scale.

There is a second advantage to following Cronbach's suggestion of comparing scale reliabilities by comparing mean within-scale interitem correlations: it allows a direct discussion of the independence of traits. As Campbell and Fiske (1959, cited by Cronbach, 1971) pointed out, two scales measuring different traits will have a lower correlation than two scales measuring the same trait. The same statement is true of mean interitem correlations within and between those scales. The mean interitem correlation tells us what reliability a scale of $\underline{n}$ items will have. If the mean correlation within a scale is sufficiently higher than the mean correlation between items on different scales, then it can be concluded that the traits assessed by those scales are different. Table 1 should help in interpreting the differences between mean interitem correlations.

Table 1

Scale Reliability (Coefficient Alpha) of a Scale Estimated from the Mean Interitem Correlation, r, and the Number of Items in the Scale, n

	Number of Items in the Scale, n									
r	3	4	5	6	7	8	9	10	11	12
.12	.29	.35	.41	.45	.49	.52	.55	.58	.60	.62
.14	.33	.39	.45	.49	.53	.57	.59	.62	.64	.66
.16	.36	.43	.49	.53	.57	.60	.63	.66	.68	.70
.18	.40	.47	.53	.57	.61	.64	.66	.69	.71	.72
.20	.43	.50	.56	.60	.64	.67	.69	.71	.73	.75
.22	.46	.53	.59	.63	.66	.69	.72	.74	.76	.77
.24	.49	.56	.61	.65	.69	.72	.74	.76	.77	.79
.26	.51	.58	.64	.68	.71	.74	.76	.78	.79	.81
.28	.54	.61	.66	.70	.73	.76	.78	.80	.81	.82
.30	.56	.63	.68	.72	.75	.77	.79	.81	.83	.84
.32	.59	.65	.70	.74	.77	.79	.81	.82	.84	.85
.34	.61	.67	.72	.76	.78	.80	.82	.84	.85	.86

Note:

Scale reliability is estimated as alpha = $nr/[1+(n-1)r]$, where r is the mean interitem correlation and n is the number of items in the scale. This estimate of reliability is the SPSS Standardized Item Alpha (Hull and Nie, 1981, p. 256).

This study used actual mean interitem correlations rather than the estimate provided by Cronbach (1951) for four reasons: (1) the actual correlations were directly available from the computer printout; (2) correlations between items on different scales were only available as actual correlations; (3) the methods typically agree within rounding error (discrepancy of mean correlation less than 0.005); and (4) use of actual mean interitem correlations allows two-item scales to be used when estimating scale reliabilities, while estimating mean correlations from alpha requires at least three items to be in each scale.

If the four scales represent distinct traits, then the items within each scale should be highly correlated and items from different scales should have somewhat lower correlations. The lowest correlations should be observed between the least similar traits. Within each booklet, a multitrait matrix based on mean observed interitem correlations can be constructed to test this assumption. Each booklet should show the same pattern of correlations since each includes items sampled from the same universe of items (at least within each age group). Accordingly, the booklets can be combined into one overall mean interitem correlation matrix. It may be optimal to do this averaging of correlation cofficients by first transforming to Fisher's z', taking the mean, and then transforming back to r. However, there is very little difference between this method

and a simple average if correlations are less than 0.50 (as judged by the close agreement between r and z' values in this range). Accordingly, simple mean values can be used to simplify calculations; the result will be an underestimation of the overall mean correlation by (probably) something less than 0.01. A second reason not to use the z' transformation is that its validity rests on the assumption of bivariate normality of the variables being correlated, an assumption that is difficult to support here.

There may be developmental differences between the age groups, but it is still reasonable to hope that the three age levels will exhibit similar patterns within their respective matrices of mean correlations. The similarities or differences between matrices should be examined with this possibility in mind.

Methodology

Correlation matrices were computed for all items in selected scales from 29 NAEP booklets over three age groups. The scales selected for analysis were the four subclassifications of NAEP Item Classification 23, with data being analyzed from Ages 9, 13, and 17. The matrices were partitioned into multitrait subcells according to the four hypothesized

traits and mean values of the observed correlations were computed for each within-trait and between-trait cell of each matrix. The mean correlation for each age group was then computed for each cell as the weighted mean of the observed booklet mean intercorrelations (weighted by the number of correlations). A "combined mean" was also computed as the unweighted mean of the three age group mean values. The result, Table 2, contains mean correlations for each subcell of the multitrait matrix for each age group and for combined ages.

After computing the three age group mean intercorrelation matrices and the one overall matrix, it was clear that the matrices were rather different for the three age groups. Those differences could be developmental, but an examination of individual item difficulties compared to correlation coefficients showed that some of the differences between age groups could have been caused by an uneven distribution of very easy or very difficult items (which had correspondingly very low interitem correlations) in the different hypothesized traits. These extreme items, along with a small number of other poorly correlated items, were then eliminated from the item pool, and the analysis was repeated. The remaining items then provided a clearer interpretation of the mean intercorrelation between and within the hypothesized traits. Summary statistics of the deleted items and correlations are

presented in Table 3 while the multitrait summary of the remaining items is presented in Table 4.

Results

Table 2 presents the results of the original multitrait matrix analysis. If these four traits are independent, then items comprising the traits should have high intercorrelation with other items in the same trait and lower intercorrelation with non-trait items. Based on this criterion, it would appear that Words and Phrases and Textual Relationship items are not independent, and that possibly Lexical Relationship items form a trait that might not be independent of Propositional Relationship items. Ages 13 and 17 appear to be comparable, although different from Age 9: Words and Phrases items seem to form a distinct trait, and possibly Lexical Relationship items do also. Once again, there is little evidence that Textual items constitute a distinct trait. The further analyses, however, indicated that these interpretations are misleading, and are the result primarily of an uneven distribution of very difficult or very easy items across the four item classifications.

While investigating the differences between the three multitrait matrices it became clear that those differences

Table 2

Mean Item intercorrelations by age level
Using all NAEP Classification 23 Items

		Words&Phr.	Lexical	Propositional	Textual
Words & Phrases	Age 9	.18 (5)	.19 (48)	.19 (79)	.16 (65)
	Age 13	.25 (14)	.21 (63)	.23 (113)	.22 (99)
	Age 17	.23 (16)	.21 (43)	.21 (65)	.20 (97)
	Combined	.22	.20	.21	.19
Lexical Rel.	Age 9		.27 (46)	.24 (149)	.21 (143)
	Age 13		.25 (36)	.23 (172)	.21 (140)
	Age 17		.25 (26)	.24 (85)	.21 (115)
	Combined		.25	.23	.21
Propos. Rel.	Age 9			.23 (158)	.22 (223)
	Age 13			.23 (182)	.22 (237)
	Age 17			.21 (54)	.19 (273)
	Combined			.22	.21
Textual Rel.	Age 9				.19 (98)
	Age 13				.21 (131)
	Age 17				.20 (115)
	Combined				.20

Notes:

Numbers in parentheses are the number of correlations upon which the mean is based.

The mean correlations for the "Combined" age groups is the unweighted mean of the three age group mean correlations.

Calculations were performed to three significant digits, but rounded to two digits for presentation here.

arose not so much because of developmental differences between the age groups but because there were different proportions of "inappropriate" items in the four trait classifications both within and across ages. Here it is necessary to be clear what constitutes an "inappropriate" item. Ideally, item difficulty will allow classification of an item on a unidimensional scale. Students with enough ability in (possession of) a trait will be able to answer correctly, and students with less than the requisite ability will succeed only by guessing. Since NAEP items have an "I don't know" foil, guessing should be minimized. An item will correlate well with other items if there is consistency in answering, if a correct answer to one item implies a correct answer to another. Guessing an answer will lead to low correlations, as does an item that almost every respondent (or almost no respondent) gets correct.

The problem with "guesswork" items in this study is that even if a trait does exist, if "inappropriate" (guesswork, easy or difficult) items comprise a large portion of the items on a scale, then a low reliability will be estimated for the scale. This is appropriate for achievement and assessment scales, but is misleading when comparing traits through mean interitem correlations (unless all scales have the same proportion of "guesswork" items).

Some of the NAEP items investigated had very low correlations with other items. For instance, item X0503D had

19 of 21 correlations with other items less than .100 while the overall mean for Age 13 Booklet 5 was r=.193. That item had a difficulty of .252, suggesting that correct responses were the result of guesswork, and the set of low correlations was the result. Most of the items with consistent low correlations could similarly be understood as items that were either too easy or too difficult. However, this was not always the case. For instance, item W0810D was very difficult (item difficulty = .189) and yet demonstrated good correlation with other items (mean r = .195) while item X1303C had item difficulty of .507 yet had 7 of 21 correlations with other items less than .100 and a mean correlation of only .122. Apparently, those who answered W0810D correctly really did know the answer while there was still a good deal of guesswork with X1303C. X1303C seems to have been one of several items that most students could answer by eliminating one or more foils and then guessing from the remaining foils.

Some of the easiest items were also anomalies. For instance X0508B had item difficulty of .945 yet the mean correlation with other items was .209; W0706B had item difficulty of .963 and mean correlation of .214. These two items apparently do accurately differentiate between those who know the item and those who do not, in contrast to item Y0405A, which had difficulty of .926 and mean interitem correlation of .099. Clearly, there is no simple way to screen items based on

item difficulty alone.

It also turns out that there was no simple way to screen items based on the response patterns. Quite comparable response patterns (percentages of respondents selecting correct and incorrect foils) emerged both with "inappropriate" items (low mean interitem correlation) and with "appropriate" items. No clear-cut algorithm for distinguishing between "appropriate" and "inappropriate" items based on item characteristics could be discerned. After all, it is possible that an item correlates poorly simply because it measures a different trait, and not because it is an "inappropriate" item. Accordingly, the decision of which items to eliminate was finally made only on the basis of multiple, low (less than .100) correlations with other items. Neither item difficulty nor item response pattern provided a consistent way to judge item value (nor did both, combined). It is recognized that this method of rejecting items is not well founded, but seemed to be the best way to proceed considering that some items should have been rejected because they were providing misleading information about the existence or nonexistence of traits.

Table 3 summarizes data about the items and booklets that were eliminated from the study. Clearly, the "inappropriate" items were not equally distributed within the four hypothesized traits. Removing the items resulted in increasing mean correlations in all cases, and could have made the existence of

Table 3

Summary statistics of deleted ("inappropriate") items and the corresponding correlations by age group

I. Deleted items.
Reported as number deleted divided by total number of items.

	Age 9	Age 13	Age 17	Combined	Pct. Deleted
Words, Phr.	2/13	1/21	1/17	4/51	7.8%
Lex. Rel.	0/28	2/31	3/22	5/81	6.2%
Prop. Rel.	1/49	3/54	3/33	7/136	5.1%
Text. Rel.	7/40	3/50	8/50	18/140	12.9%
Totals	10/130	9/156	15/122	34/408	8.3%

II. Deleted correlations.
Reported as number of low correlations deleted divided by the total number of low (less than .100) correlations.

Age 9	Age 13	Age 17	Combined	Pct. Deleted
55/61	59/65	53/54	167/180	92.8%

Note: Not all deleted correlations were low since items, rather than correlations were deleted.

III. Number of booklets affected.
Reported as number of booklets from which items were deleted divided by the total number of booklets.

Age 9	Age 13	Age 17	Combined	Pct. Deleted
5/8	6/11	9/10	20/29	69.0%

separate traits appear either more reasonable, or less so. In fact, it appears that the traits do not exist independently.

Table 4 summarizes the final set of mean interitem correlations in a multitrait matrix. Except for the Age 9 Words and Phrases mean correlations, all three age groups have quite comparable patterns of mean interitem correlations. There is some developmental change between age groups, but it is minimal and may be limited to the increase in mean item difficulty (percent of correct responses) with increasing age. Mean item difficulty (for all items) for Age 9 was .583; for Age 13 it was .739; and for Age 17 the mean item difficulty was .779. After eliminating the "inappropriate" items, mean item difficulties for the three age groups were .605, .741 and .792, respectively. It is likely that the Age 9 Words and Phrases items have different correlations than the other scale items as a result of random sampling error in the sampling of items from the universe of all possible items: there were so few Words and Phrases items involved (see Table 3).

There is an elevation of both the Words and Phrases and the Lexical Relationship item within-trait correlations compared to the other correlations, but not enough to justify identifying these categories as traits based on these data alone. It is clear that neither Propositional Relationship items nor Textual Relationship items constitutes a homogeneous scale that is distinct from the others since mean correlations

Table 4

Mean Item intercorrelations by age level
after deletion of inappropriate items.

		Words.Phr.	Lexical	Propositional	Textual
	Age 9	.22 (3)	.21 (40)	.21 (68)	.21 (41)
Words &	Age 13	.27 (11)	.22 (57)	.24 (103)	.23 (84)
Phrases	Age 17	.26 (12)	.24 (30)	.23 (51)	.22 (74)
	Combined	.25	.23	.23	.22
	Age 9		.27 (46)	.25 (147)	.24 (110)
Lexical	Age 13		.26 (33)	.24 (156)	.23 (117)
Rel.	Age 17		.26 (21)	.25 (70)	.23 (86)
	Combined		.26	.25	.23
	Age 9			.23 (147)	.23 (186)
Propos.	Age 13			.24 (169)	.24 (203)
Rel.	Age 17			.23 (45)	.22 (130)
	Combined			.23	.23
	Age 9				.21 (69)
Textual	Age 13				.23 (105)
Rel.	Age 17				.22 (82)
	Combined				.22

Notes:

Numbers in parentheses are the number of correlations upon which the mean is based.

The mean correlations for the "Combined" age groups is the unweighted mean of the three age group mean correlations.

Calculations were performed to three significant digits, but rounded to two digits for presentation here.

with items on other scales are just as high as the within-scale mean correlations. Given the reading skills context and the elevation of Words and Phrases and Lexical Relationships correlations, it may be reasonable to conclude that separate skills are involved in these two scales, even though the differences between scales are all but buried by their common elements. This judgment cannot be made on the basis of these data alone, however.

Conclusions

The data of this study alone do not support the thesis that the four subclassifications of NAEP Classification 23 are distinct traits. The apparent differences that appear when the Words and Phrases, Lexical Relationships, Propositional Relationships, and Textual Relationships items are analyzed are largely the result of unequal distribution of "inappropriate" items within the four subclassifications and not the result of existence of distinct traits. However, there is some evidence that Words and Phrases and Lexical Relationship items involve somewhat different skills, a possibility that should be examined in other substantive (rather than purely quantitative) contexts.

III. Combining Items from Different Booklets

This section of the study addresses the question, "If NAEP items from different booklets are combined into a scale, can the reliability of the resultant scale be estimated?" This problem was approached by examining interitem correlations, item difficulties, and patterns of item response. It would normally be concluded at the start that the answer to this question is "no" since no item (of those discussed here) appears in two different booklets, and since no subject completes more than one booklet. That is, the data are completely distinct. However, if the mean correlation between an item and the others measuring the same trait can be viewed as a measure of "item reliability", then it may be possible to estimate the reliability of a "new" set of items (i.e. "new" to each other). That is, the data are not really "distinct" since many items measure the same traits. If that connection to the trait can be exploited, then perhaps the reliability of a scale of "new" items can be estimated. Confidence in making this estimate will require interpretable patterns of item response as related to item difficulty and interitem correlation.

Rationale

There is no direct way to measure correlations between items, and hence no way to directly estimate the reliability of scales composed of items from different NAEP booklets (This problem will disappear in future assessments). However, these items relate to the same content area and are judged to measure the same trait; they are not independent. That lack of independence of item content will be addressed here as a potential link to allow combining items from different NAEP booklets.

Many different methods are used to estimate test reliability, including test-retest, split-half, parallel forms, Cronbach's alpha, and the standardized item alpha of Table 2. All are estimates of the same quantity: the correlation (squared) between the true score and the observed score of the scale being administered. The approach taken here is that the mean correlation between one item and the other items measuring the same trait might be interpretable as the "item reliability" for that one item.

This interpretation can be justified by thinking of different items as having both a trait component and an item-specific component. When items are combined (as when

adding vectors), the trait components add together while the item-specific components cancel. The resulting scale--if well-developed--will measure the trait exclusively, ans several such scales are "parallel" if they have the same mean, variance and covariances. If different items are parallel measures of the same trait, then the "item reliability" can be estimated as the correlation between the two items, just as scale reliability can be estimated as the correlation between parallel forms. Unfortunately, the different items do not have equal variances and covariances (This is why the "standardized item alpha" discussed earlier is not exactly equal to Cronbach's alpha). One item may have correlations with other items varying from, say, .10 to .30. The first important assumption of this method is that the mean correlation of one item with others measuring the same trait is in fact a reasonable estimate of the item reliability, even though the items are not parallel.

This assumption would be entirely reasonable if the items were parallel, or approximately so (Actually, the items need not be parallel, but only have equal variances and covariances. However, with "right-wrong" item scoring, the variance depends on the percent correct, so the two situations are equivalent). As it stands, the basis for this assumption is that the items examined are a representative sample from the set of all possible items measuring the trait, and that the

mean correlation will be interpretable as an estimate of the mean of correlations with all possible same-trait items. This assumption is comparable to (but not the same as) the fact that Cronbach's alpha is the mean of all possible split-half coefficients for a test, including those not parallel (Cronbach, 1951, p. 302) and, under certain circumstances, equals the split-half reliability coefficient for parallel forms. The comparison being made here is that mean interitem correlation is an average of "all possible pairs" just as Cronbach's alpha is the average of "all possible split halves". If the comparision is valid, then the mean interitem correlation could be interpreted as a reliability coefficient for the single item. In this way, "item reliability" can be operationally defined for an item in any NAEP booklet, based on its correlations with other items measuring the same trait.

The second important assumption of this method is that a collection of different items can be characterized by the mean of item reliabilities, and the scale reliability estimated in that way. This is the case for items measured on a single scale, so if mean item correlation does estimate item reliability, then mean item reliability should be useful for estimating scale reliability using Table 2.

In order to use this method of estimating scale reliability, it must be assumed that items examined form a fairly representative sample from the population of all

possible items. Randomness of selection would assure this representativeness, but the item selection process is not random. Nor are there enough items in every booklet to assume that a representative sample exists in all booklets, even if item selection were random. Since this is the case, a less satisfactory criterion of item acceptability was adopted: that items used be "well behaved" in that it should be possible to characterize those items by item difficulty, mean interitem correlation, and the distribution of responses to the item foils. If this cannot be done, then the relationship between items and the trait cannot be trusted enough to extrapolate from item mean correlations to scale reliabilities. This ability to characterize an item seems a reasonable expectation of items that are "well-behaved" enough to use in this analysis.

Methodology

The important part of this study is characterizing the items: How far are they from being "parallel"? How representative a sample of items exist in the NAEP booklets? How consistent and understandable are the items in terms of response patterns, item difficulties, and interitem correlations?

First, the item intercorrelation matrices were examined for each booklet. Had item intercorrelations been close enough to each other to appear that differences in correlation were the result of random sampling error, then it could be assumed that the items were close enough to "parallel" to satisfy the requirements of this study. That would require most of the correlations for an item to be within one or two standard errors of the mean correlation, with the standard error estimated as $s=(1-r^2)/SQRT(n)$, which is about .02 for the correlations and items examined here.

When it became clear that the item intercorrelations differed by far more than random sampling error would explain, then item difficulties were examined with the expectation that some or all of the correlation differences could be explained by the presence of guessing, or of floor and ceiling effects (items too difficult or too easy). If those items could be identified then they could be removed from the analysis, leaving items that were more homogeneous. For this stage of the analysis, item difficulties were compared to item intercorrelations in hopes of identifying a simple decision rule as to which items were not acceptable.

Finally, item response patterns were examined to try to relate item intercorrelations to item difficulties. Patterns of correct and incorrect responses to an item were examined for

certain items to try to characterize "guessing" in an item. The rationale here was that in some items, perhaps two foils were obviously incorrect (e.g.), and students then guessed at the correct response from the remaining foils. Such an item would be characterized by: moderate item difficulty; low interitem correlation; two foils with few respondents each; and three foils responded to in about equal proportion (with one of these being correct).

If the items can be properly classified, and items removed which appear to be untrustworthy, then estimates of scale reliabilities for scales of previously unpaired items (items from different booklets) can be made. The scale reliability might still be an inaccurate estimate, but at least a fair justification for the derivation of that reliability would exist.

Results

As discussed in Part II of this paper, no consistent patterns were found that related item difficulty to item intercorrelations and item response patterns. It was not possible to characterize items in a way that would lead to a meaningful scale reliability if that scale included items from different booklets.

Correlations for a single item could easily vary from .10 to .40, and intercorrelations for a booklet often showed an even larger spread. In no case could it be concluded that differences in item correlations were even remotely caused by random sampling error, eliminating the first, naive hope that the problem might turn out to be simple.

Some items were found to have low interitem correlations because they were too difficult or too easy. However, as discussed in Part II, there was no item difficulty upper or lower cutoff point that would allow identification of items suffering from floor or ceiling effects. For example, item W0810D had mean interitem correlation .195 and item difficulty .189 while item X0508B had mean correlation .063 and difficulty .252; item Y0405A had mean interitem correlation .099 and difficulty .926, yet item W0706B had mean correlation .214 and difficulty .963. If these items are to be characterized, it will clearly take more information than is provided by correlations and item difficulty indices alone.

The attempt to identify "guesswork" items was similarly unsuccessful. Items with different interitem correlation but the same item difficulty might have the same patterns of responses to correct and incorrect foils. Other items with comparable item response patterns might have very different interitem correlations or item difficulties. The expected patterns were found for some items, such as N0408D, which had

two foils answered very often and about equally (one of them the correct answer), yielding an item difficulty of .438 and mean interitem correlation of .133. However, too many contradictions appeared and too little consistency was found to formulate any rules about "guesswork" and related items.

It is possible that a more thorough examination could have found a pattern or a set of rules that characterize items, but such a search did not seem justified. The reasons for that decision are discussed more fully in the next section.

Conclusions

It is possible that items from different booklets could be combined under the rationale discussed earlier. Doing so with confidence would require the assumptions that mean interitem correlation is a measure of "item reliabilility" and that the items are at least somewhat understandable in terms of item difficulty, interitem correlation and item response pattern. Clearly, these items are not parallel forms of the same measure since there is such great variation in item difficulty, interitem correlation and item variance, and so the "understandability" criterion (as a practical guideline) seems needed to allow for violating the lack of common variance and covariance between items (the theoretical guideline).

It is possible that more thorough or more detailed work would allow characterizing the items in a satisfactory way, but that does not seem worth the trouble for three main reasons. First, combining items into a scale in this way is dangerous at best and could fail for reasons not being considered here (such as combining items which refer to the same text but are from different booklets--this could yield an estimated scale reliability much lower than a potentially observable one because the items relate to the same text). Putting that much effort into a project that may fail for substantive (rather than quantitative) reasons seems unwise. The second main reason for not attempting to characterize the items more closely is that there are really not very many items. Many booklets have only one or two items in a given scale, so the calculations are quite sensitive to random sampling error in the sampling of items from the set of all possible items. Even if a characterization of items is possible, then there may still be too few items to make up a meaningful scale.

Finally, it seemed undesirable to investigate the individual items further because in future NAEP analyses it will not be necessary: the balanced incomplete block (BIB) design to be used in the future (Messick, Beaton and Lord, 1983) will make this entire question unnecessary. Items from different booklets can be combined into scales of known reliability because correlations between all items will be

observed. The BIB design will actually allow a direct test of the hypothesis presented here, whether items from different NAEP booklets can be combined into a scale for which reliability can be estimated. The degree of agreement between this method and actual scale reliabilities can then be examined and this method judged thereby. Said differently, the time and energy needed to examine the items now could be better spent a few months from now in examining the method--if the method is still of interest (it may be, for analysis of earlier NAEP tapes).

Bibliography

Campbell, D.T. and Fiske, D.W. Convergent and discriminant validitation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L.J. Test validation. In Thorndike, R.L. (ed.), Educational Measurement (2nd edition). Washington, DC: American Council on Education, 1971.

National Assessment of Educational Progress, Introduction to the National Assessment of Educational Progress Public-Use Data Tapes. Denver: NAEP, 1981.

Gulliksen, H. Theory of Mental Tests. New York: Wiley, 1950.

Hull, C.H. and Nie, N.H. SPSS Update 7-9. New York: McGraw-Hill, 1981.

Johnson, N.L. and Kotz, S. Continuous Univariate Distributions--2. Boston: Houghton Mifflin, 1970.

Lord, F.M. and Novick, M.R. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968.

McConeghy, J.I. and Miller, J.D. The use of consolidated NAEP archives. A paper presented to the American Educational Research Association, Montreal, April 1983.

Messick, S., Beaton, A., and Lord, F. National Assessment of Educational Progress Reconsidered: A New Design for a New Era. Princeton: NAEP, 1983.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. and Bent, D.H., SPSS: Statistical Package for the Social Sciences (2nd ed.). New York: McGraw-Hill, 1975.

Appendix I

Mean values of NAEP variable WEIGHTS
for various booklets in three age groups

Booklet	Age 9	Age 13	Age 17
4	1401.089	1182.697	1300.167
5	1207.146	1220.267	1263.788
6	1354.850	1163.004	1234.603
7	1172.982	1233.771	1291.171
8	1247.888	1135.415	1231.500
9	1264.751	1192.719	1343.309
10	1138.221	1241.266	1284.236
11	1208.423	1266.029	1378.622
12	--	--	--
13	--	1219.883	1237.150
14	--	1227.554	1306.918
15	--	1158.455	--

Note:

These booklets are the ones containing items used elsewhere in this study. A blank entry indicates that no items from that booklet were analyzed.